REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
16-12-2010	REPRINT	

4. TITLE AND SUBTITLE
A comparative verification of forecasts from two operational solar wind models

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62601F

6. AUTHORS
Donald C. Norquist
Warner C. Meeks*

5d. PROJECT NUMBER
1010

5e. TASK NUMBER
SH

5f. WORK UNIT NUMBER
A1

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory /RVBXS
29 Randolph Road
Hanscom AFB, MA 01731-3010

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-RV-HA-TR-2010-1132

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RVBXS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; distribution unlimited.

13. SUPPLEMENTARY NOTES
Reprinted from *Space Weather*, Vol. 8, S12005, doi:10.1029/2010SW000598, 2010.
*Now at Missouri University of Science and Technology, Rolla, MO.

14. ABSTRACT
The solar wind (SW) and interplanetary magnetic field (IMF) have a significant influence on the near-earth space environment. In this study we evaluate and compare forecasts from two models that predict sW and IMF conditions: The Hakamada-Akasofu-Fry (HAF) version 2, operational at the Air Force Weather Agency, and Wang-Sheeley-Arge (WSA) version 1.6, executed routinely at the Space Weather Prediction Center. SW speed (V_{sw}) and IMF polarity (B_{pol}) forecasts at L1 were compared with Wind and Advanced Composition Explorer satellite observations. Verification statistics were computed by study year and forecast day. Results revealed that both models' mean V_{sw} are slower than observed. The HAF slow bias increases with forecast duration. WSA had lower V_{sw} forecast-observation difference (F-O) absolute means and standard deviations than HAF. HAF and WSA V_{sw} forecast standard deviations were less than observed. V_{sw} F-O mean square skill rarely exceeds that of recurrence forecasts. B_{pol} is correctly predicted 65%-85% of the time in both models. Recurrence beats the models in Ppol skill in nearly every year forecast day category. Verification by "event" (flare events ≤5 days before forecast start) and "nonevent" (no flares) forecasts showed that most HAF V_{sw} bias growth, F-O standard deviation decrease, and forecast standard deviation decrease were due to the event forecasts. Analysis of single time step V_{sw} increases of ≥20% in the nonevent forecasts indicated that both models predicted too many occurrences and missed many observed incidences. Neither model had skill above a random guess in predicting V_{sw} increase arrival time at L1.

15. SUBJECT TERMS
Forecast verification Interplanetary magnetic field Solar wind

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Donald C. Norquist
UNCL	UNCL	UNCL	UNL		19B. TELEPHONE NUMBER *(Include area code)*

SPACE WEATHER, VOL. 8, S12005, doi:10.1029/2010SW000598, 2010

A comparative verification of forecasts from two operational solar wind models

Donald C. Norquist[1] and Warner C. Meeks[1]

Received 8 June 2010; revised 4 September 2010; accepted 15 September 2010; published 16 December 2010.

[1] The solar wind (SW) and interplanetary magnetic field (IMF) have a significant influence on the near-Earth space environment. In this study we evaluate and compare forecasts from two models that predict SW and IMF conditions: the Hakamada-Akasofu-Fry (HAF) version 2, operational at the Air Force Weather Agency, and Wang-Sheeley-Arge (WSA) version 1.6, executed routinely at the Space Weather Prediction Center. SW speed (V_{sw}) and IMF polarity (B_{pol}) forecasts at L1 were compared with Wind and Advanced Composition Explorer satellite observations. Verification statistics were computed by study year and forecast day. Results revealed that both models' mean V_{sw} are slower than observed. The HAF slow bias increases with forecast duration. WSA had lower V_{sw} forecast-observation difference (F-O) absolute means and standard deviations than HAF. HAF and WSA V_{sw} forecast standard deviations were less than observed. V_{sw} F-O mean square skill rarely exceeds that of recurrence forecasts. B_{pol} is correctly predicted 65%–85% of the time in both models. Recurrence beats the models in B_{pol} skill in nearly every year forecast day category. Verification by "event" (flare events ≤5 days before forecast start) and "nonevent" (no flares) forecasts showed that most HAF V_{sw} bias growth, F-O standard deviation decrease, and forecast standard deviation decrease were due to the event forecasts. Analysis of single time step V_{sw} increases of ≥20% in the nonevent forecasts indicated that both models predicted too many occurrences and missed many observed incidences. Neither model had skill above a random guess in predicting V_{sw} increase arrival time at L1.

Citation: Norquist, D. C., and W. C. Meeks (2010), A comparative verification of forecasts from two operational solar wind models, *Space Weather*, *8*, S12005, doi:10.1029/2010SW000598.

1. Introduction

[2] The accuracy of solar wind and interplanetary magnetic field (IMF) predictions near Earth is an important factor in anticipating their effects on key space weather forecast parameters such as geomagnetic index and magnetic flux. It is useful to document the characteristics of a solar wind and IMF prediction model's performance for several reasons. Developers of the model can identify strengths and weaknesses to effectively focus their improvement efforts. Authors of other models are made aware of the formulation/deficiency relationships common to such models. Quantifying the accuracy of the model forecasting capabilities can help numerical modelers who may use the forecast output as input to drive their numerical simulations. In the operational community, forecasters gain by knowing how much confidence to place on predicted parameters. Cost/benefit information is provided to administrators who decide to sustain or replace existing models. Continuous monitoring of forecast performance can help identify changes in the measurements of instruments used to supply the reference data.

[3] Previous solar wind model forecast verification studies were designed to assess the performance of long-term simulations. *Owens et al.* [2008] used magnetic field maps constructed for entire Carrington rotations (that is, the approximately 27 day full solar rotation) to produce solar wind speed forecasts with three models. Solar wind models evaluated were the Wang-Sheeley-Arge (WSA) coronal/heliospheric model [*Arge and Pizzo*, 2000; *Arge et al.*, 2004], the WSA coronal model coupled with the ENLIL heliospheric model [*Odstrcil*, 2003], and the Magnetohydrodynamics Around a Sphere (MAS) coronal model [*Linker et al.*, 1999; *Mikić et al.*, 1999] coupled with the ENLIL heliospheric model (together called CORHEL). Forecasts from the period 1995–2002 were compared with hourly observations to compute forecast, observation, and forecast-observation difference ("error") statistics. The observations were from the Wind and Advanced Composition Explorer (ACE) satellites positioned at the L1 Lagrangian point approximately 1.5M km upstream of

[1]Battlespace Environment Division, Space Vehicles Directorate, Air Force Research Laboratory, Hanscom Air Force Base, Massachusetts, USA.

Earth. Using point-by-point statistical analysis techniques to assess the forecast performance of the models, they found that the kinematic WSA model produced solar wind speed forecasts that showed greater skill than the coupled WSA/ENLIL and CORHEL models. *Lee et al.* [2009] compared WSA/ENLIL and CORHEL solar wind forecasts using Carrington rotation magnetic field maps in 2003–2005 with ACE spacecraft measurements at L1. They showed many results from single Carrington rotations and composite histograms emphasizing the general large-scale solar wind structures from the model predictions and observations. Overall they found satisfactory agreement of the model solar wind and IMF predictions with the observed conditions. *MacNeice* [2009] ran version 1.6 of the WSA coronal/heliospheric model for Carrington rotations spanning 32 years (1977–2008), assessing the skill score and event probabilities of solar wind speed and IMF polarity predictions for each rotation. Carrington magnetic field maps were used from three solar observatories with each having separate parameter values specified in the radial velocity formula on the inner boundary. The inner boundary radius had two settings, 5 and 21.5 solar radii. Observations at L1 were used as a reference for computing forecast skill. Computing skill score with respect to persistence (actual measured value 1, 2, 4, and 8 days before the forecast valid date) as a reference, he found that WSA forecasts at L1 were competitive with persistence after 2 days of forecast time and surpass it in skill at the 4 and 8 day forecast times. The model performance was not significantly sensitive to source of the solar magnetic field data, the inner boundary radius of the WSA model, or whether the period of the solar cycle evaluated was quiet or active. WSA was better at predicting polarity reversal events than it was in forecasting high-speed events: percentage of correct forecasts (hit rate) were 61% and 40%, respectively, and percentage of incorrectly predicted event occurrence (false positive rate) of 11% and 39%, respectively.

[4] In this article we describe a study of the forecast performance of two operational solar wind models: the Hakamada-Akasofu-Fry (HAF) version 2 kinematic model currently used at the Air Force Weather Agency, and the WSA version 1.6 coronal/heliospheric model that was the version executed daily at the Space Weather Prediction Center of the National Oceanic and Atmospheric Administration as of mid-2009. We felt it was important to document the performance of these models as a baseline against which any replacement candidate model should be assessed. From the standpoint of an operational forecaster, there is interest in the day-to-day changes in the predicted state from a simulation initialized from the most recently observed conditions. Thus in this study we use the daily updated photospheric magnetic field maps as the initial conditions to drive the forecasts of the two operational models. As in the previously cited studies, we compared the forecasts of solar wind and IMF at L1 with Wind and ACE observations. We evaluated the forecast separately by forecast day in each of 6 years to investigate the dependence of model performance on forecast duration, often referred to as lead time. Forecast performance for particular years can thus be assessed as a way of determining the sensitivity of the models to solar activity level.

[5] Following this introduction, the article discusses the models and data used in this forecast verification study in section 2. Section 3 is a description of the forecast verification method. In section 4 we present the verification results from all forecasts. Separate statistics are then presented for the forecasts with and without solar disturbances in the 5 days prior to forecast initiation. This section concludes with the results of a brief study on the ability of the two models to predict single time step increases in solar wind speed. Section 5 closes the article with a summary and conclusions.

2. Data and Forecast Models

[6] The available daily magnetic field maps from the Mount Wilson Observatory (MWO) in California, referred to as the MWO Coarse Synoptic Magnetogram maps, were obtained for the odd numbered years of Solar Cycle 23: 1997, 1999, 2001, 2003, 2005, and 2007. These years were selected as the periods for model evaluation in this study as a compromise between keeping the number of forecasts executed/evaluated to a manageable size yet covering representative portions of a complete solar cycle. Each available data file represents a date during which photospheric magnetic field measurements were usually taken around local noon, weather permitting. For California, this corresponds to approximately 2000 UTC for much of the year. To account for a reasonable amount of processing time, the source surface map generated from each day's magnetogram map was used to initialize a HAF or WSA forecast beginning at 0000 UTC on the following day.

[7] The daily magnetogram map files consist of magnetic field values on a grid of 4° longitude by roughly 4.6° latitude (equally spaced in sine latitude). These grids extend longitudinally around the solar sphere between roughly ±76° latitude. The observations taken on the specified date of the file had been assimilated into the previous day's grid values as a daily update in the visible portion of the Sun. Missing data near the poles were filled in by assigning the polewardmost available value at each longitude to the missing grid points. In some cases all values along a longitude were missing. If such a data gap was ≤20° in longitude width, the magnetic field values of each latitude on the bounding longitudes were linearly interpolated to fill in the missing values of that latitude in the data gap. If the gap was greater, the magnetogram map file was not considered available for initialization of the models. The following numbers of daily magnetic field files (out of 365 possible) were available for model runs in the 6 years study, respectively: 272, 266, 248, 240, 231, and 284 for a total of 1,541 HAF and WSA forecasts.

[8] The HAF version 2 [*Fry et al.*, 2001] is a kinematic solar wind model, essentially an empirical parameterization

of plasma parcel motion, and does not contain full physical representations as does a magnetohydrodynamic (MHD) model. It runs much more quickly than an MHD model (about 6 s on a high-end workstation) and can accept rather steep spatial and temporal gradients on the inner boundary. The radial, kinematic expansion of the ejected coronal plasma and the frozen-in magnetic field defines the solar wind structure. HAF tracks the solar wind fluid parcels and the interplanetary magnetic field lines, capturing the large-scale solar wind as it flows outward from the Sun. However, the model does not provide information on the detailed energetics of the solar wind flow, nor does it resolve the small-scale waves and turbulence. HAF predicts solar wind speed, density, and magnetic field strength and orientation. In the simulation of the solar wind flow from the Sun to the Earth, radial speed is computed from the positions of the fluid parcels at successive time steps. Magnetic flux conservation is assumed for the computation of the magnetic field vector, and density is computed by assuming a conservation of mass flux. Dynamic pressure, and for events the shock arrival time (SAT) at L1, are derived from the predicted variables.

[9] HAF simulates the stream-stream interaction regions through parameterized compression-rarefaction algorithms. Projected plasma parcels encounter these interactions, arriving at the L1 point with a resulting speed that is registered as the V_{sw} prediction. HAF also simulates the plasma motion associated with nonuniform background flow resulting from solar disturbances. Flares and associated coronal mass ejections (CME) drive shock waves in the plasma that represent the leading edge of the propagating disturbance. Such a disturbance is considered an "event" for which radio, optical and X-ray observations are used to specify its characteristics in HAF. These flare properties are used as input to HAF in a list of any solar flare events that occurred in the 5 days prior to the initial time of the model execution. As explained by *Fry et al.* [2001], CME-generating disturbance events used in HAF are restricted to flares since it is difficult to specify the source information for other CME initiation mechanisms. The master list of all solar flare events used to construct each flare property input file for the HAF forecast executions is the same as that used in the three-phase "Fearless Forecast" project [*Fry et al.*, 2003; *McKenna-Lawlor et al.*, 2006; *Smith et al.*, 2009]. If no flare events occurred in the 5 days prior to the forecast initial time, the input file was blank, and the execution was considered a "nonevent" forecast. In HAF, a discontinuity in dynamic pressure (proportional to the product of density and the square of speed) represents a proxy for the shock.

[10] HAF outputs radial speed, density, magnetic field magnitude and three components of the magnetic field vector in the geocentric solar magnetospheric (GSM) coordinate system at each hour of forecast time. For an example of a time series plot of HAF forecast output, see *Fry et al.* [2001]. Because of the comparative nature of this study, only solar wind speed (V_{sw}) and IMF polarity (B_{pol}, ±1 derived from the component of the magnetic field vector along the Sun-Earth line) were considered since WSA only predicted these two quantities. A verification of all HAF output variables over the same study years was done by *Norquist* [2010]. *Fry et al.* [2003], *McKenna-Lawlor et al.* [2006], and *Smith et al.* [2009] have conducted a comprehensive evaluation of the prediction of SAT by HAF.

[11] In preprocessing for the HAF model executions, the available daily MWO magnetic field maps were interpolated to a regular 5° × 5° latitude-longitude grid. Next, the potential field source surface (PFSS) model [*Altschuler and Newkirk*, 1969] was executed on each map to extend the magnetic field from the photosphere out to 2.5 R_S in the corona also on the 5° × 5° grid. Then the velocity on the source surface at 2.5 R_S was computed by an empirical algorithm that assumes that the radial velocity on the source surface depends only on the divergence of the magnetic flux between the photosphere and the source surface. The empirical parameters in the formula are adjusted to best reproduce the observed speed at L1, and were constant for all forecasts in this study. After their ingest into HAF, the source surface magnetic field and radial velocity were interpolated to about half-degree spacing in order for HAF to produce forecasts on 1 h time steps. Each daily source surface map serves as the inner boundary condition for the 5 day HAF model execution initialized at 0000 UTC on the following day.

[12] *Arge et al.* [2004] give a detailed description of the WSA solar wind model. The design and forecast initialization procedures of the WSA have been described by *MacNeice* [2009]. Our remarks are thus limited to model aspects relevant to its execution in this study. It has inner corona, outer corona and inner heliosphere components that act in succession to propel the solar plasma from regions of open field lines in the photosphere to L1. The inner corona module of WSA uses essentially the same PFSS model as used in HAF preprocessing to extend the photosphere radial magnetic field maps out to a preliminary source surface at 2.5 R_S on a 2.5° latitude-longitude grid. The coronal extension component of the WSA model is the Schatten current sheet model [*Schatten*, 1971] that we used to compute the radial magnetic field on a source surface at 5 R_S. Next, WSA invokes the empirical scheme [*Arge et al.*, 2004] based on magnetic field expansion factors and angular distance from the nearest coronal hole boundary to compute V_{sw} on this outer source surface. The one-dimensional inner heliosphere component transported the plasma parcels radially outward with this flow speed. The parcels were accelerated or slowed by interaction with faster parcels from behind or slower parcels lying ahead as they propagated out to L1 and beyond. The magnetic field (magnitude and polarity, which is negative sunward and positive antisunward) determined on the outer source surface is similarly subject to modification due to interaction with

adjacent parcels in its transit to L1. Because of the 2.5° outer source surface grid, the WSA time step is approximately 4.55 h.

[13] Both forecast models were initialized with the same available MWO photospheric magnetic field maps to ensure direct comparability. The models were executed to predict V_{sw} and B_{pol} at their respective time step intervals over the same calendar date periods. Hourly averaged solar wind speed and magnetic field observations from the Wind (1997) and ACE (1999, 2001, 2003, 2005, and 2007) satellites at L1 served as a reference for the forecast verifications. The hourly observation times nearest the forecast valid times of the WSA time steps of each forecast day were chosen as a basis for verification. Typically, there were 26 time steps in the 5 day forecast periods whose average temporal separation from the nearest hour was approximately 15 min. Then the same hourly HAF time steps were extracted from their forecast files to represent the HAF predictions to be verified. There were periods of missing observations but they only amounted to about a 2% data loss for V_{sw} and less than 1% for B_{pol} among the hourly values used in the verifications. Occasionally (in 0.02% of the hourly outputs) HAF produced an excessive single hour value of V_{sw} that is readily apparent in the time series plots from each forecast execution. To preserve the integrity of the verification statistics, any V_{sw} prediction exceeding 150% of the maximum observed V_{sw} was not verified. Given the number of available forecasts in the 6 study years as listed above, the number of forecast time steps for verification in each forecast day ranged from a low of about 1,200 in 2005 to a high of just over 1,500 in 2007.

[14] Both HAF forecasts and ACE and Wind observations provide magnetic field vector at L1 in the three components of the GSM coordinate system. To render them directly comparable with WSA-predicted B_{pol} at L1, the x component (directed sunward from Earth) sign was noted. Positive values (sunward directed) were assigned as $B_{pol} = -1$ and negative values (antisunward) as $B_{pol} = +1$ in concert with WSA IMF polarity convention.

[15] Though the observations serve as a reference against which the forecast parameters can be evaluated, they must be qualified in their use as such a reference. There are several considerations that indicate that they are not an exact representation of the environmental truth that the model was attempting to predict. First, the forecasts represent a grid volume while observations are point measurements. This suggests a certain degree of spatial smoothing associated with the forecasts. Second, the forecasts represent a discrete time, while Wind and ACE hourly average values are the result of averaging many individual measurements over the hour following the forecast valid time (ACE Science Center, ACE data processing and archiving, 2010; available at http://www.srl.caltech.edu/ACE/ASC/docs/processing.html; see also MIT Space Plasma Group, Wind-SWE data page, 2010; available at http://web.mit.edu/space/www/wind/wind_data.html). Third, the sensors are subject to occasional solar wind and IMF disturbances, elevating parameter levels to the point where the WIND and ACE sensors can be overwhelmed and fail to produce accurate readings (ACE Science Center, ACE data processing and archiving, 2010; available at http://www.srl.caltech.edu/ACE/ASC/docs/processing.html). Fourth, the instrument sensitivity threshold dictates that very small parameter values, particularly for density, can lead to inaccurate WIND and ACE sensor measurements. For these reasons, this study refers to discrepancies between predicted and observed parameter values as differences rather than errors. Over a large sample of comparisons, as is carried out in this study, the errors in the observations tend to average out provided there is no systematic drift in the sensor. Therefore, the forecast-observation difference can be thought of as a possible model deficiency. But it must be kept in mind that some portion of that difference is due to the disparities in the source of the forecast and observed value making up the difference as mentioned above.

3. Forecast Verification Method

[16] Forecast verification by individual year allows for a look at the effect of solar activity variation on model performance. In each year's evaluation, the forecast values were verified in daylong lead time interval groups: 1–24 h, 25–48 h, 49–72 h, 73–96 h, and 97–120 h corresponding to days 1–5. Such a partitioning lets us examine forecast skill as a function of the forecast lead time. This is of interest to the operational forecaster who must assign a level of reliability to the guidance provided by each day's multiday forecast.

[17] As mentioned in section 2, forecast-observation difference represents the best measure of the forecast deficiency considering the representativeness issues. At a single model time step, this difference quantifies how much the forecast misses the mark in terms of providing a preview of future conditions at L1. The difference between mean value of the forecasts and the mean value of the observations is a systematic error called bias. If the model predicts a variable with a small mean error but a large variation in the forecast-observation difference, the model performs poorly in matching the temporal variation of the observed state. So both the systematic and random components of the forecast-observation difference tell a story about the nature of the model's forecast performance. In this study we sought to examine both components to most fully evaluate each model's predictive ability.

[18] For solar wind speed we were able to compute a full range of forecast-observation difference (F-O) statistics. Representing the F-O of a parameter x at each hour i by

$$X_i = x_{F_i} - x_{O_i},$$

where F and O designate the forecast and observed values, respectively, the following F-O statistics were computed:
mean,

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i;$$

mean square,

$$\overline{X^2} = \frac{1}{N}\sum_{i=1}^{N} X_i^2;$$

standard deviation,

$$\sigma_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2};$$

absolute mean,

$$|\overline{X}| = \frac{1}{N}\sum_{i=1}^{N}|X_i|;$$

skill score,

$$SS = 1 - \overline{X^2}/\overline{P^2};\ P_i = x_{O_{i-6}} - x_{O_i},$$

in which P_i is the persistence F-O, subtracting the observed value nearest each forecast time step from the observed value at the initial time (0 h) of the 5 day forecast. This initial observed value is held constant ("persisted") over the entire forecast period and serves as the forecast value at all time steps in this simplistic type of forecast. In all statistical quantities, the summation is over N forecast time steps in each daylong lead time interval group over all available forecasts in a study year. In the skill score, the persistence forecast serves as a baseline. A skill score value of 1 (variable difference mean square = 0) means that the forecast matched the observations perfectly. With a skill score of 0 the evaluated model predicts the observations no better than persistence, while negative values denote poorer agreement with observations than persistence.

[19] An alternative baseline forecast against which a model prediction can be compared in skill score is recurrence. Because the Sun rotates with a period of approximately 27 days, operational forecasters often take advantage of the long-lasting nature of some major solar features (e.g., coronal holes) in making space weather predictions. The state of forecast parameters of interest 27 days prior to the forecast valid date are commonly used as a "first guess" for the formulation of a forecast. Recognizing the fact that the solar period is not exactly 27 days and that, while general features may remain their spatial detail can change significantly during Sun's revolution, we used the daily averaged observations from 27 days prior as the recurrence forecast. *Norquist* [2010] showed that the daily average recurrence consistently demonstrates superior skill over hourly recurrence (27 days prior to the hour nearest the time step valid time). Skill scores based on day average recurrence were also computed in the HAF and WSA V_{sw} forecast verification.

[20] Correlation of the forecast values with the observed values is a quantitative measure of how well the temporal variations match. It can be computed for both the time step value pairs and the day average value pairs. In either case, the correlation is computed using the respective means and standard deviations of the forecast and observed values, then using them to compute the correlation:
mean,

$$\overline{x}_F = \frac{1}{N}\sum_{i=1}^{N} x_{F_i},\ \overline{x}_O = \frac{1}{N}\sum_{i=1}^{N} x_{O_i};$$

standard deviation,

$$\sigma_{x_F} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{F_i} - \overline{x}_F)^2},\ \sigma_{x_O} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{O_i} - \overline{x}_O)^2};$$

correlation,

$$r_{x_F x_O} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_{F_i} - \overline{x}_F)(x_{O_i} - \overline{x}_O)}{\sigma_{x_F}\sigma_{x_O}},$$

where the x_F, x_O symbols can represent either the hourly or day average value of the respective forecast or observed values within a forecast interval and the overbars signify their averages over all time step values in a forecast interval in a given year or over all day average values for a forecast interval in a given year. Not surprisingly, because of the smoothing resulting from the day averaging, the day average values generally result in a somewhat higher correlation than the individual time step values.

[21] Because B_{pol} is a binary variable, we limited our assessment to the percentage of correct polarity forecast time steps in each forecast day category. By extension, the mean square of F-O, divided by four to represent a unitary difference, was computed and thus a skill score determined. This was done for both a persistence and day average (of 27 day prior hourly B_{pol} observations) recurrence baseline.

[22] In addition to the verification of all forecasts as a whole, a separate assessment was made of the HAF and WSA forecasts in disturbed and quiet solar conditions. The full set of forecasts was segregated into event and nonevent categories as described in section 2. That is, any forecast for which the HAF flare input file had any flare characteristics specified for any of the 5 days prior to the initial time of the forecast was considered an "event" forecast period for both models. Forecast initial conditions without 5 day prior flares were deemed "nonevent" forecasts. All of the statistics described above were computed separately for event and nonevent forecasts from both models. In addition to assessing model performance in disturbed and quiet solar periods, this breakdown allows an investigation of the

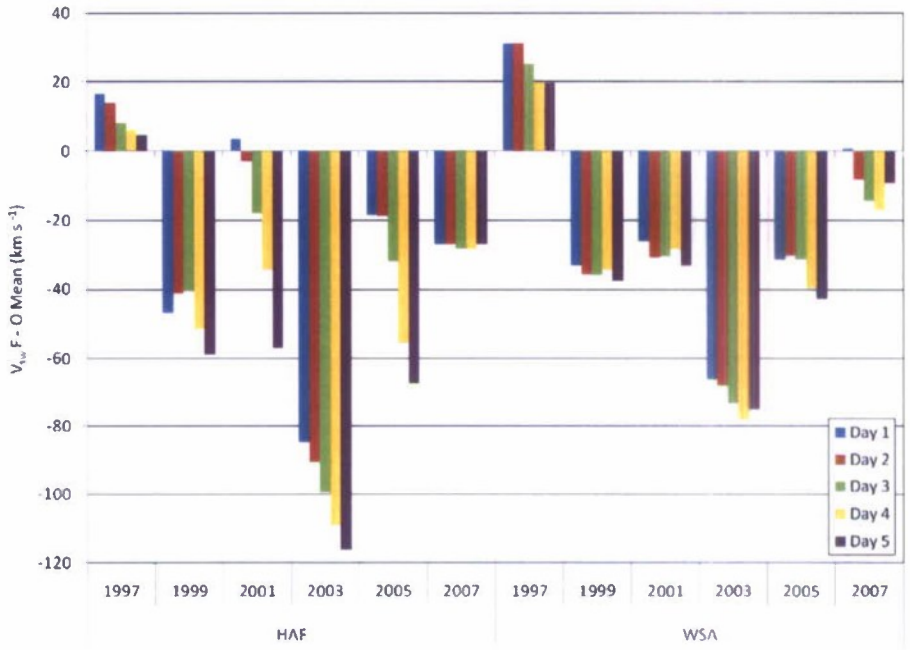

Figure 1. Solar wind speed (V_{sw}) forecast-observation difference (F-O) mean (km s^{-1}) for all HAF and WSA forecasts in each study year, computed separately for forecast days 1–5.

effect of attempting CME simulation on the HAF solar wind speed and IMF polarity predictions.

[23] Because of the geomagnetic storming implications of high-speed events (HSEs) arriving at Earth, it is of interest to evaluate the ability of the models to predict them. *MacNeice* [2009] included an evaluation of HSEs predicted by WSA in his study. As mentioned previously, the "Fearless Forecast" project evaluated shock arrival at L1 by HAF and other models in which sudden increases in dynamic pressure represented a shock. In the current study we used V_{sw} increases in single WSA time step intervals to indicate the occurrence of HSEs. We found that a search with a threshold of a single time step V_{sw} increase of ≥20% detected roughly the same number of HSEs as shocks identified in HAF forecasts in the same study years by *Norquist* [2010]. Searching the HAF, WSA, and observed 5 day forecast periods for all nonevent forecasts, to avoid the CME simulations in HAF which are not available in WSA forecasts, yielded counts of HSEs yes/no predicted and observed for which contingency tables could be constructed. Common skill scores were computed from the contingency tables for both HAF and WSA forecast periods.

4. Results

4.1. Forecast-Observation Difference Statistics

[24] We first examine the systematic error of the V_{sw} forecasts as reflected in the forecast-observation differences (F-O) mean. To set the context, the observational mean solar wind speed (V_{sw}) and magnitude of the magnetic field vector |B| for the 6 study years are displayed in Table 1. In Solar Cycle 23 solar activity with regard to sunspot number reached a peak in mid-2000 and stayed high until beginning to decline in early 2002 (National Geophysical Data Center, ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_NUMBERS/AMERICAN/, see SMOOTHED.PLT for smoothed monthly mean sunspot number). Both 1997 and 2007 were solar minimum years, 1999 was in the ascending phase and 2003 and 2005 were in the declining phase. To minimize systematic error the forecast models must reproduce the climatology of the solar wind so that the annual F-O mean computed as described in section 3 for each forecast day is small. They are shown in Figure 1 for the HAF and WSA forecasts. In five of the years both models display a negative bias relative to the observed mean. Even with the greater observed mean in 2003, as a percentage of the observed mean the F-O mean is largest in 2003 for both models, as great as −21% for HAF and −14% for WSA. Another notable property of the F-O mean is that the negative bias of the HAF forecast mean V_{sw} increases with increasing forecast lead time in five of the 6 years. This is only apparent in 3 years for WSA and to a much lesser extent. The observed mean (not shown) remains virtually unchanged with forecast lead time as expected. The speed bias over all years and forecast days was −34 and −22 km s^{-1} for

Table 1. Annual Mean of Available Hourly Averaged Observations of Solar Wind Speed (V_{sw}) and of the Magnitude of the Interplanetary Magnetic Field Vector (|B|) as Measured at the L1 Lagrange Point by Wind (1997) and ACE (Other Years) Sensors for the 6 Study Years

	1997	1999	2001	2003	2005	2007
V_{sw} (km s^{-1})	381	448	439	549	486	490
|B| (nT)	6	7	7	8	6	5

HAF and WSA, respectively, representing −8% and −5% of the overall observed mean V_{sw}.

[25] The magnitude of the forecast departure from the mean is represented by the F-O absolute mean. This is shown as a percentage of the observed mean V_{sw} in Figure 2. By this metric, a measure of the ability of the models to replicate the observations, HAF had its worse performance in 2007 while WSA displayed somewhat smaller F-O in its worst years of 2001 and 2005. There is no indication of the growing slow bias of V_{sw} with lead time for HAF in this graph. This indicates that it was due to an increase in the number of negative F-O time steps with greater forecast lead time rather than the difference magnitude growing larger at those time steps. This was confirmed with histograms of F-O counts by discrete F-O size bins (not shown) showing an increasing number of counts of the negative difference categories with increasing forecast day. V_{sw} F-O magnitude in the 6 study years as a percentage of the observed mean V_{sw} is 20.3% for HAF and 17.8% for WSA.

[26] Next we consider the random component of the F-O, represented by the standard deviation of the forecast-observation differences about their mean shown in Figure 3. The F-O standard deviation is a metric of how greatly the forecast-observation difference varies. If the forecast time series tracked with the temporal variation of the observations, the random component of the error would be zero and only a simple tuning of the model to correct the systematic bias would be needed. Generally, model simulations are less temporally variable, or smoother, than nature so much of the difference standard deviation can be explained by the inability of the model to capture all of the natural variance. Another prominent cause is timing errors, e.g., if a high-speed enhancement was predicted to arrive at L1 too early or too late. In Figure 3 we see that HAF exaggerates the F-O standard deviation early in the forecast period in 2001 and 2005, to a lesser degree in 1997, 1999, and 2003, and then significantly damps the variance in days 4 and 5. WSA F-O standard deviation shows no clear forecast lead time trends, and identifies 2001 and 2005 as the years with the poorest simulation of the observed V_{sw} variability. Though for most forecast day-year categories HAF displays larger F-O standard deviation than WSA, day 5 values for HAF are less than WSA's in 1997, 1999, and 2001. The F-O standard deviation over all time steps

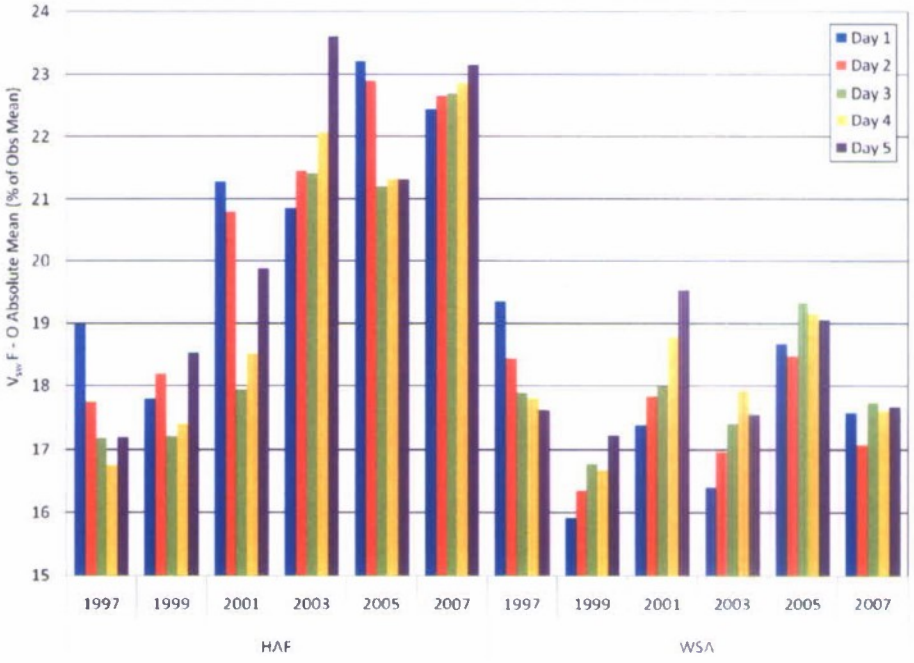

Figure 2. Same as Figure 1 except for F-O absolute mean as a percentage of the annual mean of observations valid on each forecast day.

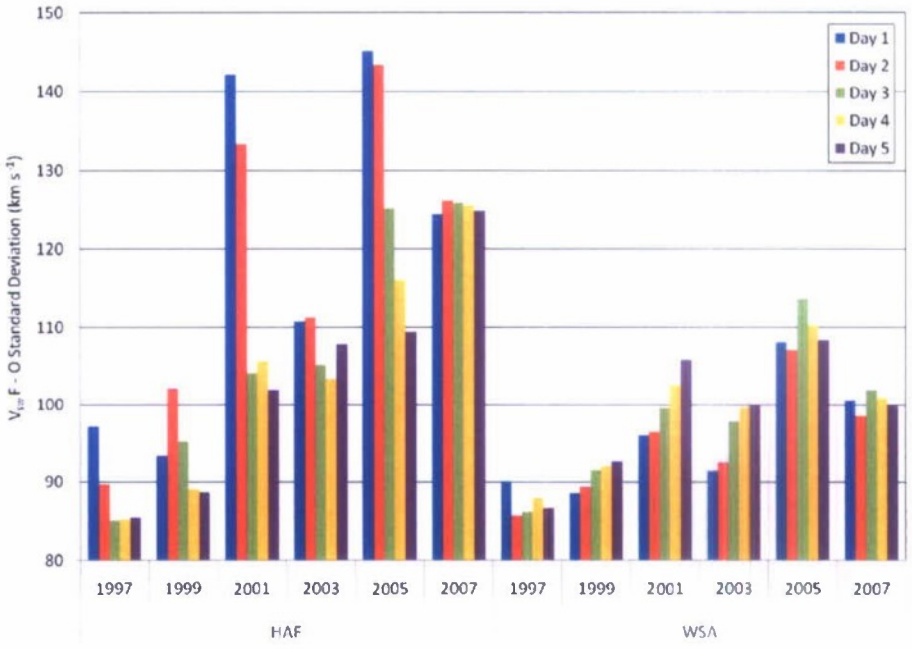

Figure 3. Same as Figure 1 except for F-O standard deviation.

in all years was 111.3 and 97.5 km s^{-1} in the HAF and WSA V_{sw} forecasts, respectively.

[27] It was mentioned in the previous paragraph that models usually underrepresent the temporal variance of a prognostic quantity. In the case of V_{sw}, this was examined by evaluating the full standard deviation of V_{sw} as observed and as predicted by HAF and WSA. When the standard deviation was assessed separately by year and by forecast day, the latter showed a significant decline from day 1 to day 5 in the HAF forecasts. Figure 4 indicates that HAF forecasts begin with a standard deviation similar to that of observations, which is then severely damped below even WSA's that is consistently less than observed. This result is consistent with the excessive F-O standard deviations early in HAF forecasts in Figure 3. That is, when standard deviations of both HAF and observations are large, the difference standard deviations are likely to be large as well. The overall V_{sw} standard deviation for HAF, WSA, and observations is 88.1, 78.9, and 98.8 km s^{-1}, respectively. The latter two values compare with 84.3 for WSA and 99.2 km s^{-1} for observations as computed by *Owens et al.* [2008] for 1995–2002.

[28] The day average forecast and observed means and standard deviations were used to compute the day average forecast versus observation correlations shown in Figure 5. The forecast day average correlations are slightly higher than their single time step counterparts in all forecast day-year categories for both models. Figure 5 indicates that there is significant variation in the correlations among the study years, from values less than 0.2 in 2001 to values above 0.5 in 2003 for both models. The year-to-year change is the same for the models until 2007, when the HAF correlations are smallest and WSA's are second to largest. This is consistent with the relative values of F-O standard deviation in 2007 as seen in Figure 3. Over all years and forecast days, the day-average V_{sw} correlations are 0.32 and 0.42 for HAF and WSA, respectively.

[29] We now turn our attention to skill score of the V_{sw} forecasts. Table 2 presents the skill score values with respect to persistence and recurrence for all forecast day-year categories. The results indicate that, except for early in the forecast period (day 1 and for the first 4 years day 2), recurrence is a higher standard as a reference for model forecasts than is persistence. Persistence is expected to beat either model or recurrence in the first forecast day since on many days V_{sw} changes slowly. However, recurrence excels over persistence in skill (i.e., the models have a more negative score with respect to recurrence) by day 2 or 3. In fact, as the values in Table 2 indicate, recurrence is superior in skill (i.e., show a negative skill score) to the models on days 4 and 5 in all years for HAF and 3 years for WSA, whereas both models beat persistence on those days in all but one forecast day-year category. Except for 1997, WSA skill scores are greater than HAF's with respect to both references in all of the forecast day-year categories. In regards to recurrence skill score, it is clear from Table 2 that neither model shows any discernable V_{sw} prediction improvement or degradation trend with forecast lead time. The WSA persistence-based skill scores computed by *MacNeice* [2009] for MWO initial

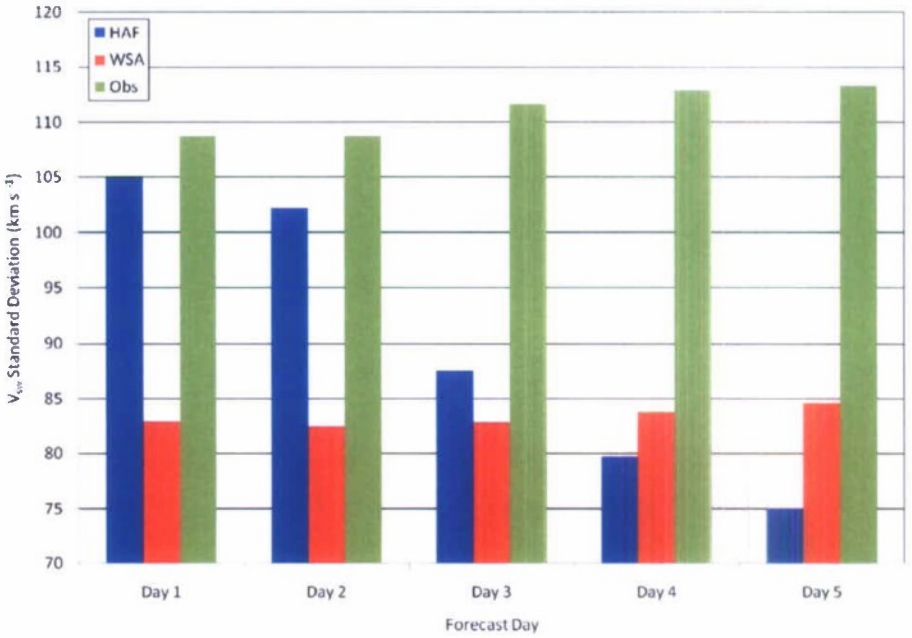

Figure 4. Standard deviation of HAF and WSA V_{sw} forecasts and observations (Obs) computed over all study years by forecast day.

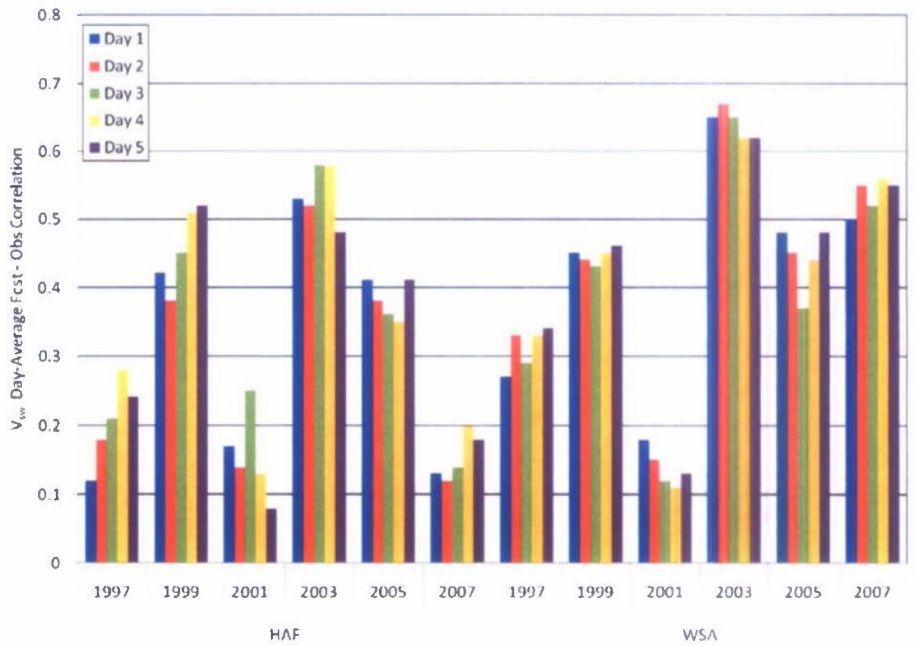

Figure 5. Correlations of day average HAF and WSA V_{sw} forecasts with day average observations in each study year, computed separately for forecast days 1–5.

Table 2. V_{sw} Skill Scores With Respect to Persistence (Per) and Recurrence (Rec) for All Forecast Day-Year Categories in This Study

	HAF					WSA				
	Day 1	Day 2	Day 3	Day 4	Day 5	Day 1	Day 2	Day 3	Day 4	Day 5
1997										
Per	−5.57	−0.89	−0.09	0.09	0.12	−5.15	−0.91	−0.20	−0.01	0.05
Rec	−0.43	−0.24	−0.09	−0.13	−0.18	−0.34	−0.26	−0.21	−0.25	−0.28
1999										
Per	−2.90	−0.47	0.10	0.27	0.29	−2.19	−0.12	0.19	0.33	0.37
Rec	−0.03	−0.13	0.01	−0.01	−0.10	0.16	0.14	0.11	0.07	0.03
2001										
Per	−5.30	−0.88	0.09	0.09	0.11	−2.07	−0.09	0.12	0.17	0.20
Rec	−0.90	−0.65	0.00	−0.11	−0.24	0.07	0.05	0.03	−0.02	−0.11
2003										
Per	−3.89	−0.56	0.00	0.12	0.12	−2.21	0.00	0.29	0.38	0.45
Rec	−0.33	−0.46	−0.39	−0.38	−0.45	0.12	0.07	0.01	0.02	0.10
2005										
Per	−4.26	−0.73	0.16	0.39	0.47	−2.12	−0.02	0.30	0.50	0.56
Rec	−0.75	−0.73	−0.23	−0.17	−0.06	−0.04	−0.02	−0.03	0.03	0.12
2007										
Per	−5.03	−0.47	0.13	0.31	0.36	−2.75	0.14	0.45	0.56	0.61
Rec	−0.92	−1.01	−1.03	−1.10	−1.10	−0.20	−0.18	−0.29	−0.32	−0.30

conditions and 5.0 R_s source surface were −1.19, −0.16, and 0.18 for days 1, 2, and 4. In comparison, the corresponding WSA skill scores computed in this study for the same forecast days over all study years were −2.5, −0.07, and 0.39. The significantly better performance of day 1 persistence over WSA in the current study may be due to using the observed value at the 0 h of the forecast throughout the first day, rather than the 24 h earlier value at each forecast time step as was done by *MacNeice* [2009].

[30] The comparative performance of the two models in predicting B_{pol} at L1 is first displayed as the percentage of correct forecasts as shown in Figure 6. The pattern of

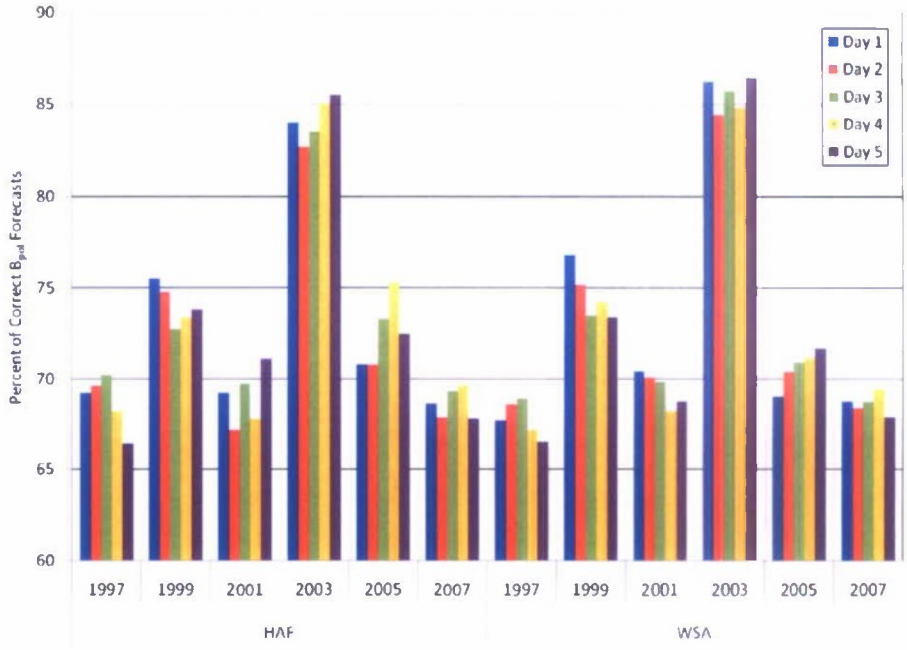

Figure 6. Percent of correct magnetic field polarity (B_{pol}) forecasts by study year and forecast duration.

Table 3. B_{pol} Skill Scores With Respect to Persistence (Per) and Recurrence (Rec) for All Forecast Day-Year Categories in This Study

	HAF					WSA				
	Day 1	Day 2	Day 3	Day 4	Day 5	Day 1	Day 2	Day 3	Day 4	Day 5
1997										
Per	0.07	0.24	0.29	0.25	0.27	0.02	0.21	0.26	0.23	0.27
Rec	−0.14	−0.15	−0.11	−0.13	−0.19	−0.19	−0.19	−0.16	−0.17	−0.18
1999										
Per	−0.08	0.21	0.30	0.34	0.42	−0.02	0.23	0.32	0.36	0.42
Rec	−0.01	−0.08	−0.07	−0.10	−0.12	0.04	−0.06	−0.04	−0.07	−0.14
2001										
Per	−0.15	0.11	0.23	0.25	0.36	−0.10	0.19	0.23	0.26	0.30
Rec	−0.17	−0.26	−0.14	−0.15	−0.14	−0.12	−0.15	−0.13	−0.14	−0.23
2003										
Per	0.01	0.29	0.45	0.59	0.66	0.15	0.36	0.53	0.59	0.69
Rec	−0.21	−0.21	−0.17	−0.07	−0.07	−0.05	−0.09	−0.01	−0.09	0.00
2005										
Per	−0.16	0.15	0.36	0.48	0.48	−0.23	0.14	0.31	0.40	0.46
Rec	−0.25	−0.22	−0.20	−0.08	−0.25	−0.32	−0.23	−0.30	−0.25	−0.29
2007										
Per	0.02	0.17	0.30	0.33	0.34	0.03	0.19	0.29	0.32	0.34
Rec	−0.32	−0.41	−0.35	−0.39	−0.37	−0.32	−0.38	−0.37	−0.39	−0.37

performance by year is same for both models, and the actual values are similar, too. In all years and forecast days, the percent of correct B_{pol} forecasts is virtually the same for HAF and WSA, 72.7% and 72.6%, respectively. It is interesting to note that in the year with the largest IMF magnitude, 2003, the percentage of correct forecasts by both models is 10%–15% greater than the other years. *Norquist* [2010] also found that skill in predicting the magnetic field vector azimuth angle (the angle the vector makes with the Sun-Earth line when projected on the ecliptic plane) was much better predicted in 2003 by HAF than for the other years. *MacNeice* [2009] found that the same configuration of WSA using MWO magnetograms produced B_{pol} predictions that match observations in 76% of the forecast times.

[31] The final comparative metric for the B_{pol} forecasts is skill score. Both persistence and day average recurrence were used as the reference as was done for the V_{sw} forecasts. Table 3 shows the skill score values for all forecast year categories. As was the case for V_{sw}, recurrence forecasts are tougher for the models to beat than persistence. But in B_{pol}, this begins with day 1 as evidenced by the larger negative values at all 5 days in a majority of the years for both models. As was seen in the percent of correct B_{pol} forecasts in Figure 6, 2003 has the best B_{pol} skill scores from both models with respect to persistence. However, there is no clear-cut best year in regards to recurrence. In fact, neither model exceeds recurrence in skill, only in a single forecast day-year category (1999-1) in the WSA predictions is the model better. Nor does either model show any clear trend of forecast skill as a function of forecast lead time with respect to recurrence. In fairness to the models, the recurrence B_{pol} "forecast" value was computed as the average of the hourly observed −1 and +1 values 27 days prior to the valid time forecast day. As such the average could have any value between the two polarity values and thus would result in a lower F-O mean square than would result if a −1 or +1 value was imposed as the recurrence forecast. Persistence also did not have that advantage, as a single value of B_{pol} (the 0 h observation) was used. The mean square F-O was computed over all forecast days and study years and was, from best to worst: Rec, 0.23; HAF, 0.27; WSA, 0.27; and Per, 0.49. In other words, the B_{pol} forecasts from HAF and WSA excelled over persistence but were inferior to recurrence. The WSA persistence-based skill scores for B_{pol} computed in this study were somewhat better than those of *MacNeice* [2009]: day 1, 2, and 4 values were −0.03, 0.21 and 0.35 in the current study and −0.83, 0.01, and 0.04 according to his evaluation.

4.2. Event Versus Nonevent Forecasts

[32] HAF and WSA forecasts were also assessed separately for event and nonevent 5 day forecast periods. In event forecasts, HAF inputs included flare properties for at least one flare event in the 5 days prior to forecast period start. These enabled HAF to simulate a CME propagating to L1 in accordance with the assumptions detailed by *Fry et al.* [2001]. In nonevent cases both models operated in the absence of such disturbances. The number of WSA forecast time steps, at which both models were compared with observations in the respective conditions, are shown in Figure 7.

[33] Space does not allow us to reproduce all of the F-O statistics charts of section 4.1 separated by event and nonevent forecasts. So we show the more telling aspects of the effects of disturbed versus quiet conditions on the forecasts of the two models. We begin with the standard

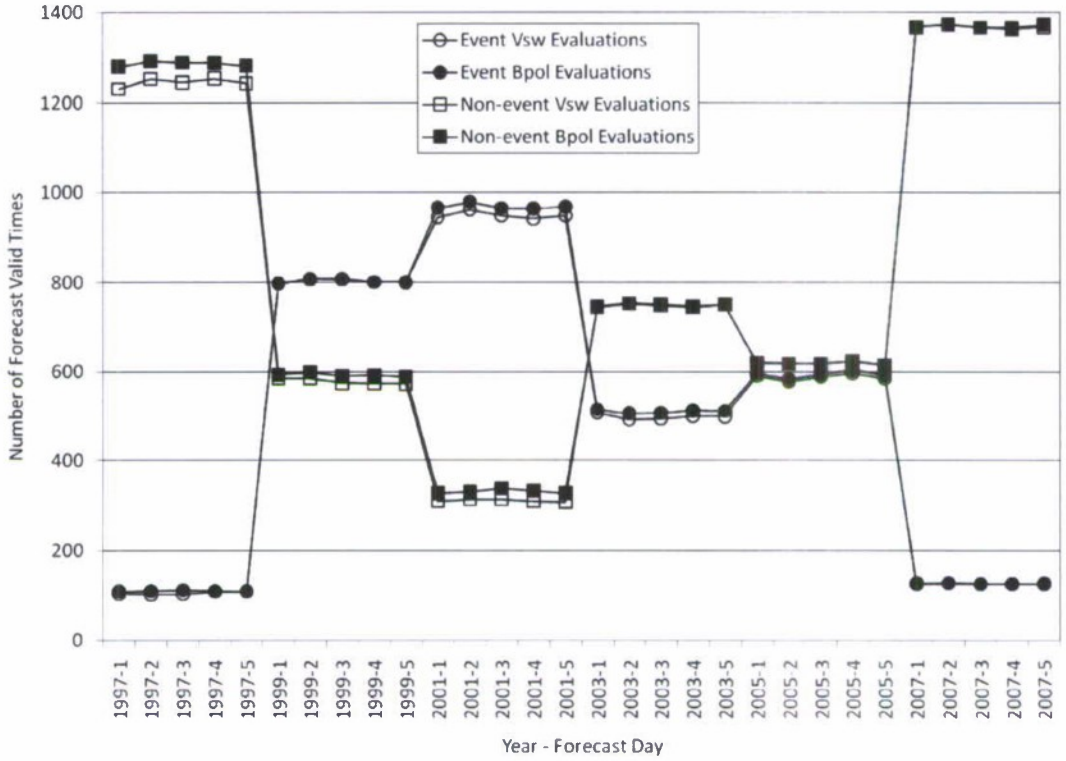

Figure 7. Number of forecast time steps used in the verification of event and nonevent forecasts for each forecast day and year for both V_{sw} and B_{pol} forecasts.

deviation of the V_{sw} forecasts and observations as shown in Figure 8. These results make it clear that all of the dramatic drop in standard deviation of HAF V_{sw} forecasts seen in Figure 4 is due to the event forecasts. In the first 2 days of predictions HAF V_{sw} standard deviation exceeds even those observed. They drop down to a level that is replicated in all 5 days of the nonevent forecasts. The latter is surprisingly similar to that of the WSA forecasts in disturbed conditions but lower than the standard deviation of WSA forecasts in quiet conditions. As is seen in Figure 4, both models' variance is well below that of the observations in either of the conditions after the severe damping in the HAF event forecasts.

[34] As one would suspect from this result, almost all of the F-O mean decrease with forecast day in HAF predictions (Figure 1) was due to the event forecasts (not shown). This was also true of the HAF V_{sw} F-O standard deviations from event forecasts (not shown) in which the days 1 and 2 values are greater than 160 km s^{-1} in 2005 (compare with Figure 3 for all forecasts). The profile and magnitude of the WSA nonevent V_{sw} F-O standard deviations look very much like those of Figure 3 for all forecasts while for the event forecasts the years at and after solar maximum (2001–2005) have the largest values. WSA event forecasts had somewhat larger F-O standard deviations than did the nonevent cases unlike the forecast V_{sw} standard deviations in Figure 8. The WSA event and nonevent V_{sw} skill scores were very much alike.

[35] In regards to B_{pol} prediction performance in event and nonevent forecasts, we found that they were better in nonevent periods in all study years but one. This is seen in Figure 9, which shows the percent of correct B_{pol} forecasts over all forecast days for each study year. Uncertainty is greatest in event forecasts of 1997 and 2007 due to the relatively few forecast time steps used in the verification as seen in Figure 7. In agreement with the results over both disturbed and quiet conditions shown in Figure 6, 2003 was the best year for both models in both conditions, due perhaps in part to the greater number of nonevent than event verification time steps (Figure 7). By contrast, in 2001 when event forecasts dominate, there are fewer correct B_{pol} forecasts (Figure 6) and the contrast between event and nonevent skill is greatest (Figure 9).

4.3. High-Speed Event Analysis

[36] We examined both models' forecasts and the observations in all of the nonevent 5 day forecast periods for single (4.55 h) time step V_{sw} increases of 20%. In Figure 10 we show a forecast period beginning on 29 April 1997 at 0000 UTC in which all three had such a

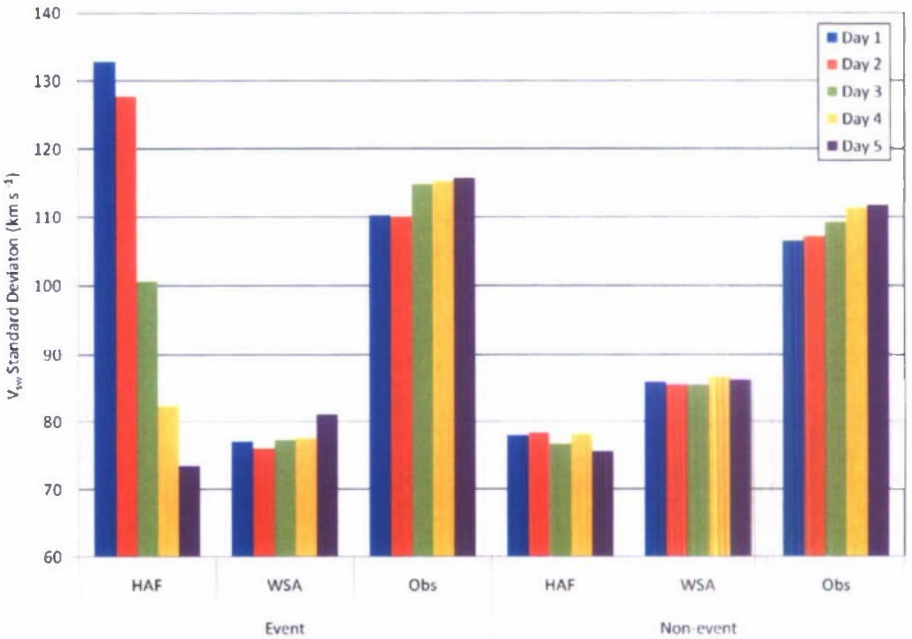

Figure 8. Standard deviation of HAF, WSA, and observed V_{sw} computed separately for event and nonevent forecasts over all study years by forecast day.

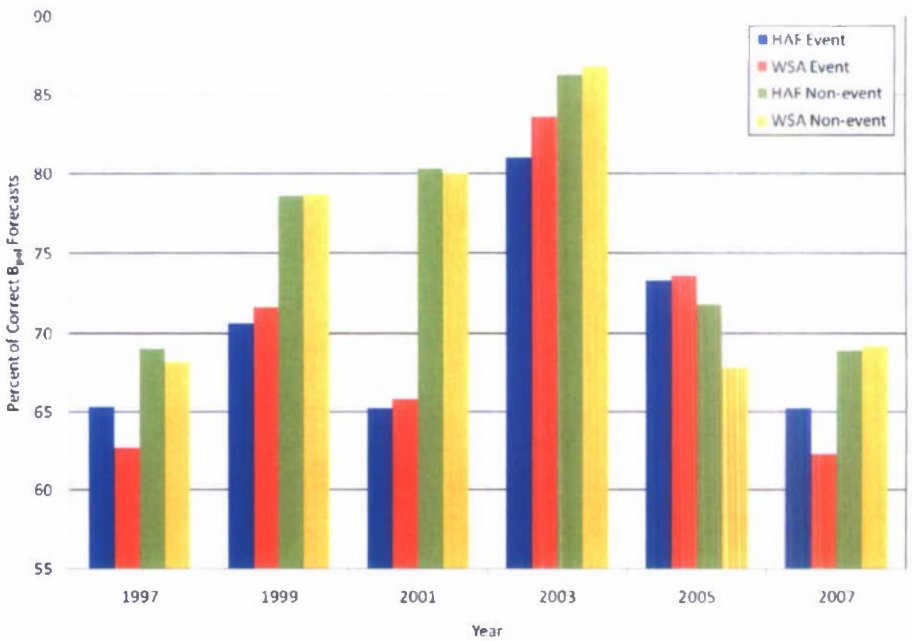

Figure 9. Percent of correct B_{pol} forecasts for HAF and WSA event and nonevent forecast periods over all forecast days by study year.

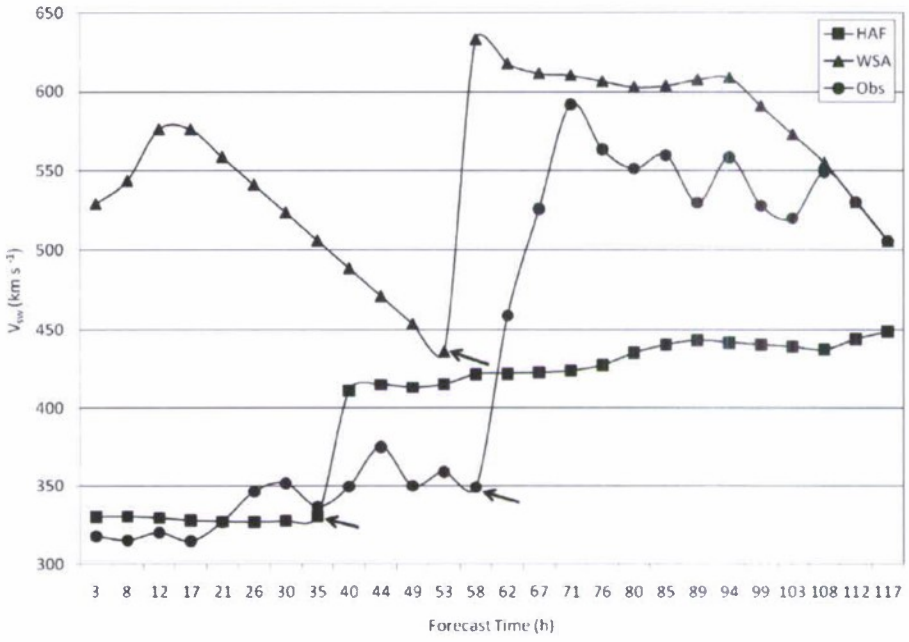

Figure 10. Time series of HAF and WSA predicted and observed V_{sw} for the nonevent forecast initialized 0000 UTC 29 April 1997. The arrows denote the time step of the HSEs.

high-speed event (HSE). Of the 1,534 complete (i.e., 120 h) HAF and WSA forecasts, 935 were nonevent with forecast and observation V_{sw} values available for at least half of the time steps. They were assessed and counted by 5 day forecast periods with and without at least one qualifying HSE. Contingency tables listing these counts for HAF versus Obs and WSA versus Obs are given in Table 4.

[37] Quantitatively, we can assign comparative skill of the contingency tables using metrics computed from the table entries. Setting a = F-yes/O-yes, b = F-yes/O-no, c = F-no/O-yes, d = F-no/O-no, and n = a + b + c + d, *Wilks* [1995] defines: hit rate (HR) = (a + d)/n, critical success index (CSI) = a/(a + b + c), probability of detection (POD) = a/(a + c), false alarm rate (FAR) = b/(a + b), and bias = (a + b)/(a + c). Values of these metrics are presented in Table 5.

Table 4. Contingency Table Counts of Forecast/Observed 5 Day Nonevent Forecast Periods With (Yes) and Without (No) at Least One HSE as Defined in the Text

	Forecast	
Observed	Yes	No
	HAF	
Yes	82	164
No	205	484
	WSA	
Yes	127	120
No	246	442

[38] The HR is a fraction of the forecast periods that were correctly predicted in regards to yes/no HSE occurrence. According to the HSE criteria used in this study, both models anticipated the correct outcome in 61% of the nonevent forecast periods. CSI is the ratio of correctly predicted HSE forecasts to the total that were forecast and/or observed, a measure of ability to anticipate HSEs while avoiding misses and false alarms. It omits the null events (F-no/O-no) and puts more stress on correct occurrences. By this measure both models perform quite poorly. Many of the observed HSEs were missed by the forecasts, yet they both predicted an excessive number of HSE occurrences that did not happen. The former shortcoming is quantified in the POD and the latter in the FAR. Bias is simply a ratio of the total number of predicted HSE periods to observed HSE periods. WSA overpredicted the number of periods with HSEs by 51% compared to 17% by HAF. Yet the POD metric indicates that HAF only predicted 33% of the forecast periods with at least one observed HSE while WSA predicted 51% of them. WSA predicted 30% more HSE forecast periods than did HAF.

Table 5. Quantitative Skill Metrics Computed From Contingency Table Counts in Table 4[a]

	HR	CSI	POD	FAR	Bias
HAF	0.61	0.18	0.33	0.71	1.17
WSA	0.61	0.26	0.51	0.66	1.51
Perfect	1.0	1.0	1.0	0.0	1.0

[a]See text for definition of metrics.

[39] By way of perspective, if two coins were flipped together 1,000 times and the number of heads/tails combinations was counted, there would be approximately the same number (roughly 250) of each of the four combinations. In this case, the metric scores would be HR = 0.5, CSI = 0.25, POD = 0.5, FAR = 0.5, and Bias = 1. The HAF and WSA forecasts have a slightly better HR but no better CSI, POD, FAR and Bias than flipping a pair of coins.

[40] Evaluating the HSE arrival time (HAT) at L1 for the F-yes/O-yes cases, we find the following outcomes. HAF had a forecast-observed HAT difference mean, absolute mean and root-mean-square of −6.5, 26.9, and 33.1 h for its 82 periods. The corresponding values for WSA's 127 periods were −0.5, 25.6, and 32.7 h. Selecting a HAT for the 82 HAF periods at random yielded values of these metrics of −3.2, 25.5, and 29.1. Thus, in the limited number of forecast periods for which the models correctly predicted that an HSE would occur, the predicted time of its arrival at L1 was no better than a random guess.

5. Summary and Conclusions

[41] Two solar wind models used routinely, the Hakamada-Akasofu-Fry (HAF) version 2 and the Wang-Sheeley-Arge (WSA) version 1.6, were evaluated through verification of daily 5 day forecasts on dates with available solar magnetic field maps over 6 years of Solar Cycle 23. The two prognostic variables of the WSA, radial solar wind speed (V_{sw}) and the attendant frozen-in magnetic field polarity (B_{pol}), as predicted by both models at the L1 Lagrange point near Earth out to 120 h were compared with in situ observations from the Wind and ACE sensor suites at that location. First, a number of statistical quantities based on the V_{sw} forecast-observation differences (F-O) at the WSA model time steps (and using the nearest hourly time step of HAF) were computed to quantify the models' performance in predicting solar wind speed. In addition, the percent of correct ±1 B_{pol} predictions were assessed along with the B_{pol} skill score. In computing skill score for both V_{sw} and B_{pol}, we used a persistence forecast (the 0 h observation of the 5 day forecast period) and a recurrence prediction (the 27 day prior day average of hourly observations) as the reference. Second, we split the forecast periods into event and nonevent forecasts based on the presence or absence of documented flare events in the 5 days prior to forecast initial time. Verification statistics were computed separately for the two conditions to highlight the effects of disturbed and quiet solar conditions on the forecast behavior of the models. Third, the ability of the models to correctly predict occurrences of high-speed events (HSEs) represented by single time step V_{sw} increases of 20% or more was evaluated. The nonevent 5 day forecast periods were examined for the prescribed HSEs and counts of periods with and without HSEs in both the forecast and observations were conducted to produce a contingency table for both models. Skill statistics were computed for both models, as well as forecast-observation difference statistics for the HSE arrival time at L1 for forecast-yes/observed-yes cases.

[42] Results revealed that both models' V_{sw} are on average slower than observed, by as much as 21% for HAF and 14% for WSA. The HAF slow bias increases with forecast lead time, while no such trend was apparent in WSA forecasts. V_{sw} mean forecast-observation difference (F-O) magnitude in the 6 study years as a percentage of the observed mean V_{sw} is 20.2% for HAF and 17.8% for WSA. The fact that it did not increase with time of HAF forecasts, along with evidence from F-O histograms, made it clear that the increase in slow bias was due to a growing number of negative F-O time steps and not an increase in their magnitude. Overall V_{sw} F-O standard deviation for WSA is 13.8% less than for HAF. In HAF forecasts it decreases 13.1% from day 1 to day 5 while WSA forecasts gain 3% over the 5 days. HAF and WSA V_{sw} forecast standard deviation is 10.7% and 19.9% less than observed, respectively, suggesting that both models underrepresent the temporal variability of the solar wind speed. In three of the 6 study years it dropped significantly (by over 50% in two of the years) during the 5 day HAF forecasts while staying steady in WSA predictions. Correlations of the day average forecast and observed V_{sw} were 0.32 for HAF and 0.42 for WSA, reflecting a similar limited ability to match the temporal variation of the observations. Even the year-to-year trend of correlations was alike except for the last study year. Computation of V_{sw} skill score with respect to persistence and recurrence showed that, except for the first forecast day, recurrence beats persistence in a mean squared difference from observations. V_{sw} F-O mean square skill exceeded that of recurrence forecasts in only one (HAF) and 15 (WSA) of the 30 year forecast day categories. B_{pol} at L1 was correctly predicted in 65%–75% of the time steps in both HAF and WSA forecasts, and as high as 85% when the IMF strength is greatest. Overall B_{pol} forecast accuracy was the same for both models (73%). In B_{pol}, recurrence beat HAF in all and WSA in all but 1 year forecast day category. It is also consistently a better forecast than persistence. Notably, neither HAF nor WSA show any trend in V_{sw} or B_{pol} prediction skill score with forecast lead time when recurrence is used as a baseline.

[43] When the forecasts were separated into "event" (when flare properties were specified for HAF forecasts) and "nonevent" (no flares prior to HAF forecast) cases, we found that almost all of the HAF V_{sw} negative bias increase, F-O standard deviation decrease, and forecast standard deviation decrease with forecast lead time were due to the event forecasts. There was little sign of such trends in the HAF nonevent forecasts of V_{sw}. This suggests an impact of the simulation of transients on the solar wind flow in the HAF forecasts in two ways: it causes excessive variability early in the forecast and retards the plasma flow as the forecast progresses. WSA V_{sw} F-O standard deviations were larger for event than nonevent forecasts in the years at and after solar maximum. In neither condition

did they show any trend with forecast lead time as did HAF in V_{sw} event forecasts. Both models demonstrated somewhat higher B_{pol} prediction skill in nonevent than event forecasts.

[44] Single model time step increases of 20% or more in V_{sw}, the criterion used to denote HSEs, were analyzed in the 5 day nonevent forecast periods for both models and observations. Both models produced an excessive number of forecasts with HSEs. More of the forecast periods with observed HSEs were missed by HAF than were predicted, while WSA predicted about half of them. Neither model demonstrated any skill above a random guess in regards to predicting the HSE arrival time at L1 in forecast periods with predicted and observed HSEs. These results suggest that there is a lot of room for improvement in the prediction of high-speed streams and corotating interaction regions.

[45] In summary, the WSA model performed somewhat better in V_{sw} prediction than HAF, whereas they were about even in B_{pol} skill. HAF V_{sw} event forecasts were subject to decreasing speed throughout the integration and excessive variance earlier in the forecast period that was damped below that of HAF by day 5. In quiet solar conditions both models underrepresent the temporal variability of the observed V_{sw}. Recurrence still remains a better forecast than what can be produced by the models especially in magnetic field polarity. These findings accentuate the challenges involved in the realistic simulation of the solar wind and its attendant magnetic field. Future evaluations of more advanced physics models should shed light on how their performance varies with forecast lead time and solar activity level.

[46] Acknowledgments. We thank Ghee Fry of Exploration Physics, Inc. for his guidance in setting up and executing the HAF model, as well as for providing the event files listing the specification of flare properties. We express our appreciation to our colleague Nick Arge for providing the WSA model code, the MWO magnetic field map files, and the Wind/ACE observations. The WIND satellite data were originally obtained from Massachusetts Institute of Technology Space Plasma Group and National Aeronautics and Space Administration, and the ACE data were obtained from the ACE Science Center at California Institute of Technology. Funding for the second author was provided by the Space Vehicles Directorate Space Scholars program. Overall funding and support for this project was provided by the applied research program and the Space Weather Forecasting Laboratory of the Air Force Research Laboratory.

References

Altschuler, M. D., and G. Newkirk (1969), Magnetic fields and the structure of the solar corona. I: Methods of calculating coronal fields, *Sol. Phys.*, *9*, 131–149, doi:10.1007/BF00145734.

Arge, C. N., and V. J. Pizzo (2000), Improvements in the prediction of solar wind conditions using near-real time solar magnetic field updates, *J. Geophys. Res.*, *105*, 10,465–10,480, doi:10.1029/1999JA000262.

Arge, C. N., J. G. Luhmann, D. Odstrcil, C. J. Schrijver, and Y. Li (2004), Stream structure and coronal sources of the solar wind during the May 12th, 1997 CME, *J. Atmos. Sol. Terr. Phys.*, *66*, 1295–1309, doi:10.1016/j.jastp.2004.03.018.

Fry, C. D., W. Sun, C. S. Deehr, M. Dryer, Z. Smith, S.-I. Akasofu, M. Tokumaru, and M. Kojima (2001), Improvements to the HAF solar wind model for space weather predictions, *J. Geophys. Res.*, *106*, 20,985–21,001, doi:10.1029/2000JA000220.

Fry, C. D., M. Dryer, Z. Smith, W. Sun, C. S. Deehr, and S.-I. Akasofu (2003), Forecasting solar wind structures and shock arrival times using an ensemble of models, *J. Geophys. Res.*, *108*(A2), 1070, doi:10.1029/2002JA009474.

Lee, C. O., J. G. Luhmann, D. Odstrcil, P. J. MacNeice, I. de Pater, P. Riley, and C. N. Arge (2009), The solar wind at 1 AU during the declining phase of Solar Cycle 23: Comparison of 3D numerical model results with observations, *Sol. Phys.*, *254*, 155–183, doi:10.1007/s11207-008-9280-y.

Linker, J. A., Z. Mikić, D. A. Biesecker, R. J. Forsyth, S. E. Gibson, A. J. Lazurus, A. Lecinski, P. Riley, A. Szabo, and B. J. Thompson (1999), Magnetohydrodynamic modeling of the solar corona during Whole Sun Month, *J. Geophys. Res.*, *104*, 9809–9830, doi:10.1029/1998JA900159.

MacNeice, P. (2009), Validation of community models: 2. Development of a baseline using the Wang-Sheeley-Arge model, *Space Weather*, *7*, S12002, doi:10.1029/2009SW000489.

McKenna-Lawlor, S. M. P., M. Dryer, M. D. Kartalev, Z. Smith, C. D. Fry, W. Sun, C. S. Deehr, K. Kecskemety, and K. Kudela (2006), Near real-time predictions of the arrival at Earth of flare-related shocks during Solar Cycle 23, *J. Geophys. Res.*, *111*, A11103, doi:10.1029/2005JA011162.

Mikić, Z., J. A. Linker, D. D. Schnack, R. Lionello, and A. Tarditi (1999), Magnetohydrodynamic modeling of the global solar corona, *Phys. Plasmas*, *6*, 2217–2224, doi:10.1063/1.873474.

Norquist, D. C. (2010), Verification of forecasts from the Hakamada-Akasofu-Fry v2 solar wind model, *Rep. AFRL-RV-HA-TR-2010-1010*, 69 pp., Air Force Res. Lab., Air Force Mater. Command, Hanscom Air Force Base, Mass.

Odstrcil, D. (2003), Modeling 3-D solar wind structures, *Adv. Space Res.*, *32*, 497–506, doi:10.1016/S0273-1177(03)00332-6.

Owens, M. J., H. E. Spence, S. McGregor, W. J. Hughes, J. M. Quinn, C. N. Arge, P. Riley, J. Linker, and D. Odstrcil (2008), Metrics for solar wind prediction models: Comparison of empirical, hybrid, and physics-based schemes with 8 years of L1 observations, *Space Weather*, *6*, S08001, doi:10.1029/2007SW000380.

Schatten, K. H. (1971), Current sheet magnetic model for the solar corona, *Cosmic Electrodyn.*, *2*, 232–245.

Smith, Z. K., M. Dryer, S. M. P. McKenna-Lawlor, C. D. Fry, C. S. Deehr, and W. Sun (2009), Operational validation of HAFv2's predictions of interplanetary shock arrivals at Earth: Declining phase of Solar Cycle 23, *J. Geophys. Res.*, *114*, A05106, doi:10.1029/2008JA013836.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences*, 467 pp., Academic, San Diego, Calif.

W. C. Meeks and D. C. Norquist, Battlespace Environment Division, Space Vehicles Directorate, Air Force Research Laboratory, Hanscom Air Force Base, MA 01731, USA. (afrl.rvb.pa@hanscom.af.mil)